COMPARATIVE ANALYSIS OF TEXRANK AND LEXRANK MODELS FOR TEXT SUMMARIZATION: ARCHITECTURE, PERFORMANCE, AND PROSPECT

A THESIS SUBMITTED TO

THE FACULTY OF ARCHITECTURE AND ENGINEERING

OF

EPOKA UNIVERSITY

BY

David Keci

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR BACHELOR DEGREE

IN COMPUTER ENGINEERING

MONTH (MAY), Year (2024)

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last name: David Keci

Signature:

# **ABSTRACT**

COMPARATIVE ANALYSIS OF TEXRANK AND LEXRANK MODELS FOR TEXT SUMMARIZATION: ARCHITECTURE, PERFORMANCE, AND PROSPECT

[David Keçi](mailto:dkeci21@epoka.edu.al)

M.Sc., Department of Computer Engineering

Supervisor: Msc.Igli Draci

When it comes to condensing the most important information from a document, text summarizing is an indispensable tool that may be used in a variety of contexts, including academic research, news, and internal documentation. Selecting between extractive and abstractive methods, as well as accurately capturing text meaning and assessing summarizing quality, is one of the main issues in text summarization. Moreover, the achievement of ideal summaries depends on the suitable tool selection. This study compares and contrasts the TexRank and LexRank algorithms, looking at their approaches, output, and text summarizing capabilities. Through examining these models, the study seeks to offer discernments into the development and efficacy of text summarization methodologies, facilitating decision-making for relevant parties in various sectors and use cases.

**Keywords:** text summarization,TextRank,LexRank, algorithms.

***Dedicated to………….…***

# 

**TABLE OF CONTENTS**

[ABSTRACT](#_gjdgxs)

[CHAPTER 1](#_1t3h5sf)

Introduction

1.1 Text Summarization: Overview and Challenges

1.2 Text summarization as part of NLP

1.3 Introduction to TextRank and LexRank Models

CHAPTER 2

Models and Literature Review

2.1 Explanation of TextRank and LexRank Models

2.2 Previous Studies on TextRank Model

2.3 Previous Studies on LexRank Model

CHAPTER 3

Methodology

3.1 Data Collection

3.2 Experimental Design

3.3 Evaluation Metrics

Comparative Analysis

4.1 Architecture of LexRankModel

4.2 Architecture of TexRank Model

4.3 Performance Comparison

4.4 Prospects and Challenges

CHAPTER 4

Results and Discussion

5.1 Comparative Analysis Results

5.2 Discussion on Key Findings

[CONCLUSION](https://docs.google.com/document/d/1fU5k2z95zJwBucFxXrdJvJ1j5ktBFK3OedFj1iG8xzw/edit#heading=h.23ckvvd)

[REFERENCES](https://docs.google.com/document/d/1fU5k2z95zJwBucFxXrdJvJ1j5ktBFK3OedFj1iG8xzw/edit#heading=h.ihv636)

[APPENDIX A](https://docs.google.com/document/d/1fU5k2z95zJwBucFxXrdJvJ1j5ktBFK3OedFj1iG8xzw/edit#heading=h.1hmsyys)

[Ex. CALIBRATION CURVE FOR DETERMINATION OF CELL NUMBER](https://docs.google.com/document/d/1fU5k2z95zJwBucFxXrdJvJ1j5ktBFK3OedFj1iG8xzw/edit#heading=h.41mghml)

# **ABSTRACT**

Two graph-based methods, LexRank and TextRank, are introduced for computing the relative importance of textual units in Natural Language Processing (NLP). LexRank focuses on sentence salience in Text Summarization (TS) by computing sentence importance based on eigenvector centrality in a graph representation. This model utilizes a connectivity matrix derived from intra-sentence cosine similarity. The system employing LexRank ranked first in multiple tasks in the DUC 2004 evaluation, demonstrating its effectiveness. Results indicate that degree-based methods, such as LexRank, outperform centroid-based methods and other systems in most cases. Additionally, the LexRank with threshold method demonstrates superior performance over other degree-based techniques. TextRank, another graph-based ranking model, is introduced for text processing. Its successful application in various natural language applications is showcased.

# **CHAPTER 1**

# **INTRODUCTION**

***1.1 Text Summarization: Overview and Challenges***

Summarization of textual data refers to the automatic process of producing a short and compressed version of a given text while preserving the information carefully selected for its users. Summarization is vital for any information retrieval and management endeavor since it helps the user understand the essence of a long document or distributed corpus in a short time. An extractive summarizer chooses sentences directly from the original text to include in the output summary. This contrasts with an abstractive summarizer that rephrases and synthesizes new sentences based on concepts in the original. Despite the increasing attention that abstractive summarization continues to receive, extractive summarization processes in several works better because of its data-driven nature. The summarization also faces challenges choosing between abstractive and extractive summarizers, semantic gist, performance measures, and deciding on which tool to use among others.

***1.2 Text summarization as part of NLP***

Text summarization is one of many submodules of Natural Language Processing (NLP), which is aimed at enabling computers to understand and generate human language . The classes of graph-based ranking algorithms developed for tasks such as citation summary ,web-page ranking, TexRank, LexRank, Seq2Seq and others can be applied to NLP text summarization . These algorithms are known to make local ranking choices by aggregating information from all over the text and find numerous applications, including automated key phrase extraction, text summarization, and word sense disambiguation . One of NLP’s foundational concepts is powerful statistical models, which have been successfully applied to diverse problems areas, such as parsing, machine meaning decoding, and automatic summarizing.

***1.3 Introduction to TextRank and LexRank Models***

Kleinberg’s HITS algorithm, and Google’s PageRank are two well-known graph-based ranking algorithms, and they have been used in many fields, including NLP. These methods rank vertices in a graph based on global information generated from the entire graph. For the lexical or semantic graphs of natural language documents, similar approaches can be applied to generate graph-based ranking models for each NLP task. TextRank and LexRank models for graphs were subsequently developed using these ranking principles. and have introduced several important works based on graph ranking; for example, unsupervised keyword and sentence extraction have become feasible. In automated keyphrase and sentence extraction exercises, they have outperformed existing transformational systems.

# **CHAPTER 2**

# **Models and Literature Review**

***2.1.1TextRank Model***

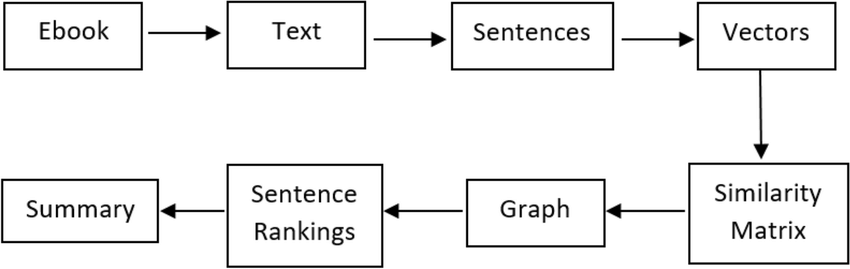
The TextRank method is an unsupervised approach based on graph theory that calculates a sentence weight or the measure of the relevance of the given sentence relative to the whole text. It has become popular recently due to the study of Mihalcea who described the textual analog of Google's PageRank algorithm. In addition, the program generates graphs of natural language texts based on their structural connections and ranks these phrases according to their inclusions in a given context.

When it comes to the actual operation of TextRank it perceives each sentence as a node in a network and it compares the content overlap of the nodes by measuring the extent to which each node is unique to others. Subsequently, sharing edges between vertices is displayed when they have a signiﬁcant amount of similarities between them. The TextRank technique extracts sentences that are the most relevant to the given content by somehow navigating this network to extract the data that has been encrypted in the text.

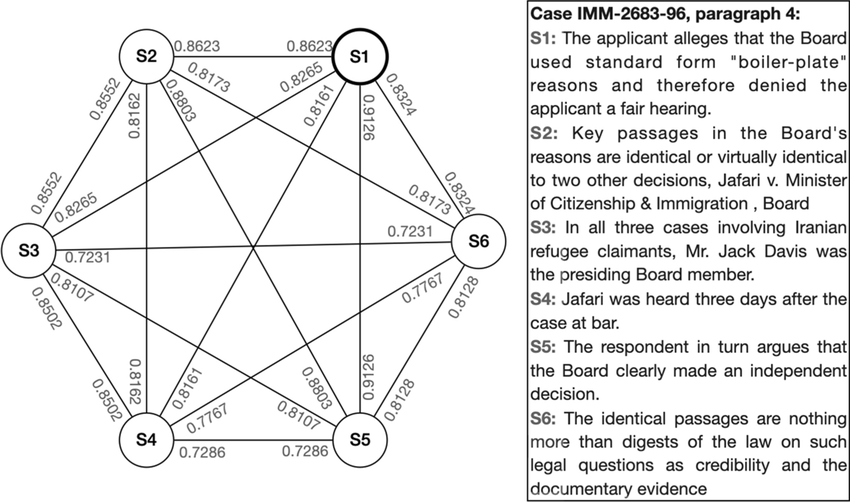
In consideration, however, the global information extracted from the entire graph edges through the different algorithms of graph-based ranking is the main reason why the TextRank algorithm is used to emanate the sentence's relevance. Staking a node is a way of voting cast in the network when the connections are created, similar to "endorsing" other vertices in a manner resembling voting mechanism. This significance depends on the number of vertices (or users), which are relevant, an endpoint will endorse on top of the endorsements they receive themselves.

TextRank uses some damping coefficient that gives a score for each node in an algorithmic manner. This score shows the possibility of creating an edge between the two vertices with each subsequent use of this edge by the graph-traversing. The scores are going through this procedure again and again until they become very close to the threshold of the predetermined threshold. In doing so, the system cross checks any sentence with others in order to distinguish the most important ones that are directly related in the text.

Taking this a step forward, alongside apparently just scoring a passage, TextRank usefulness includes the capability to create nickname summaries by just picking sentences that have highest scores. TextRank is PageRank algorithm based but expandable into a graph model so that other networks like Positional function and HITS(Hyperlink-Induced Topic Search) can be incorporated. By its nature of the ability to process various forms of a text it becomes a more valuable tool than before. (Smys et al. #), (Gonzalez)



Picture 2.1

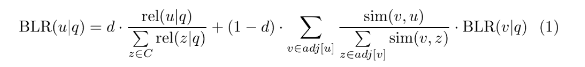


Picture 2.2 (“An example of TextRank graph, each sentence in the paragraph is...”)

***2.1.2 LexRank Model***

LexRank, a cool method of stochastic graph-based notation, makes things a lot easier in finding the meaning of textual units in summarization (TS). LexRank extends on conventional techniques built upon the keywords or prototypes for sentences, using the transformation from sentences into a graph with eigenvector centrality as support. With this novel initialization, LexRank establishes a graph in which each sentence is mapped to a node and edges are created between nodes based on sentence comparability achieved through cosine-similarity between the sentences.In practice, LexRank calculates the value of each sentence through eigenvector centrality, which depends on its degree and the importance of connected nodes. This process entails subsequent iterations, similar to walking to neighboring nodes, to determine the overall scores.

The performance of this technique is very much verified as it took the first position in DUC 2004 evaluation which is one of the most prestigious evaluations, thereby, fortifying its capability of producing very high quality summaries. In the comparative studies, focus is laid on degree-based methods such as LexRank, showing that they outperform centroid based methods, and other systems competing in the evaluation. Even with noise from imperfect clustering, LexRank’s performance is further elevated by applying thresholding. This supports a claim that even data noise produced from wrong topical clustering, does not affect LexRank’s ability to perform. The LexRank indeed stands as a foundational building block in the text-mining methodologies offering a trustworthy and qualitative means of having information of importance from text data. LexRank is a pioneering model designed to tackle the challenge of Question Answering. It draws its mathematical inspiration from computing the stationary distribution in DMCM. In Biased LexRank (B-LR), a graph is constructed where each node represents a question-answer (QA) pair, and edges are weighted by the similarity of text between two QA pairs. The order of the stationary distribution for QA pairs serves as the rank for output.

The main formula used is 

where d is the damper factor determining the probability of restarting the run-dom walk. For the link weight, they use cosine similarity computed through tf and idf scores to represent the strength of support between two QA pairs. (Du et al. #) (Erkan and Radev)

2.2 Previous Studies on TextRank Model

2.3 Previous Studies on LexRank Model

**References**

**Chapter 2**

**2.1.1**

* Smys, S., et al., editors. *Second International Conference on Computer Networks and Communication Technologies: ICCNCT 2019*. Springer International Publishing, 2020. Accessed 7 May 2024.
* Gonzalez, Ruddy. “TextRank: Bringing Order into Texts - Rada Mihalcea and Paul Tarau Department of Computer Science University of North Texas.” *Electrical Engineering and Computer Science*, https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf. Accessed 7 May 2024.
* “An example of TextRank graph, each sentence in the paragraph is...” *ResearchGate*, https://www.researchgate.net/figure/An-example-of-TextRank-graph-each-sentence-in-the-paragraph-is-represented-as-a-vertice\_fig4\_362616496. Accessed 7 May 2024.

**2.1.2**

* Du, Xiaoyong, et al., editors. *Web Technologies and Applications: 13th Asia-Pacific Web Conference, APWeb 2011, Beijing, Chiina, April 18-20, 2011. Proceedings*. Springer, 2011. Accessed 7 May 2024.
* Erkan, Gunes, and Dragomir R. Radev. “[1109.2128] LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.” *arXiv*, 9 September 2011, https://arxiv.org/abs/1109.2128. Accessed 7 May 2024.